Large-Scale Adaptive Semi-Supervised Learning via Unified Inductive and Transductive Model 1-KDD 2014

Active-Transductive Learning with Label-Adapted Kernels 2-KDD 2014

Semisupervised Dimensionality Reduction and Classification Through Virtual Label Regression 3-IEEE Transactions on ...Part B (Cybernetics) 2011

Transductive Learning on Adaptive Graphs 4-AAAI 2010

Transductive Classification via Local Learning Regularization 5-AISTAS 2007

Linear Manifold Regularization with Adaptive Graph for Semi-supervised Dimensionality Reduction 6-IJCAI 2017

Semi-supervised manifold regularization with adaptive graph construction 7-PRL 2017

2018.4.19

Active Learning via Transductive Experimental Design 8-ICML 2006

Non-greedy Active Learning for Text Categorization using Convex Transductive Experimental Design 9-SIGIR 2008

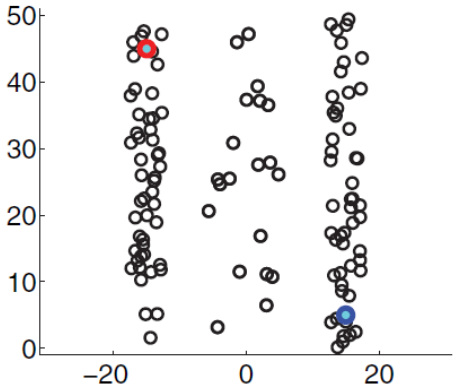
Beyond the Point Cloud: from Transductive to Semi-supervised Learning 10-ICML 2005

Manifold Regularized Experimental Design for Active Learning 11-TIP 2017

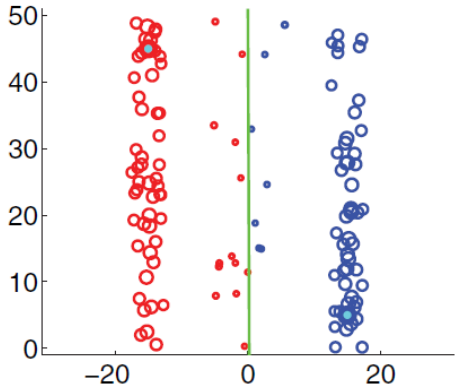
Diversifying Convex Transductive Experimental Design for Active Learning 12-IJCAI 2016

2018.4.19

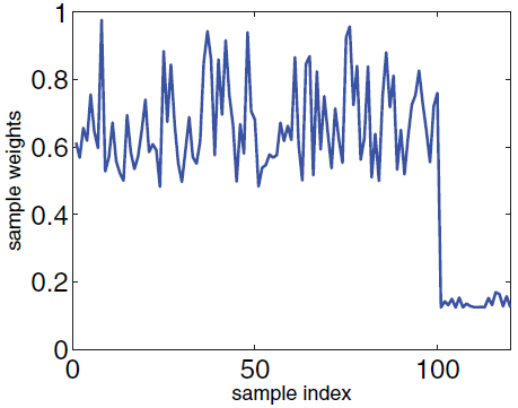
1-KDD 2014



(a) Original toy data



(b) Toy data after classification using our model

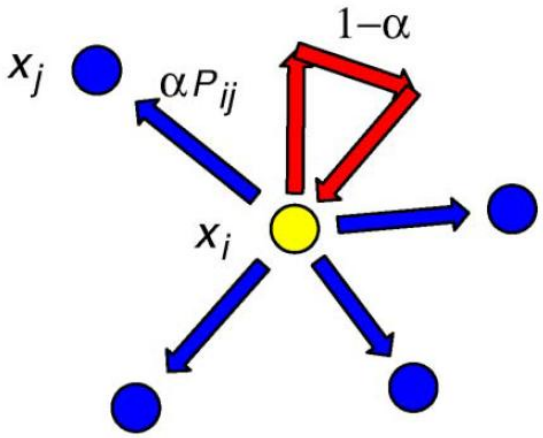


(c) Sample weights

3-TCYB 2011

In order to capture the outliers or the data from the class other than the labeled classes, we consider special random walks

$$\mathbf{A}_{ij} = e^{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}/\sigma^2 \qquad \mathbf{P} = \mathbf{D}^{-1}\mathbf{A} \qquad \mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$$



$$\tilde{\mathbf{P}} = \mathbf{I}_\beta + \mathbf{I}_\alpha \mathbf{P} \qquad \mathbf{I}_\beta = \mathbf{I} - \mathbf{I}_\alpha$$

$$\mathbf{Y} = [\mathbf{Y}_1^T, \ldots, \mathbf{Y}_n^T]^T \in \mathbb{R}^{n \times (c+1)}$$

add an additional class $c + 1$in order to detect outlier data

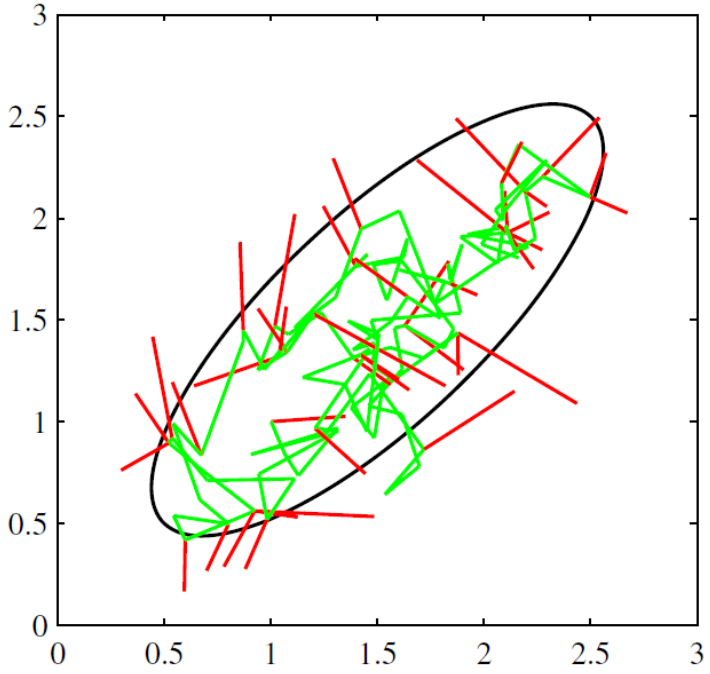
$$\mathbf{G} = \mathbf{I}_\beta + \hat{\mathbf{P}}\mathbf{I}_\beta + \hat{\mathbf{P}}^2\mathbf{I}_\beta + \cdots + \hat{\mathbf{P}}^n\mathbf{I}_\beta + \cdots . \qquad \hat{\mathbf{P}} = \mathbf{I}_\alpha \mathbf{P}$$

$$\mathbf{G} = (\mathbf{I} - \mathbf{I}_\alpha \mathbf{P})^{-1}\mathbf{I}_\beta$$

the probability of the i_{th} point stopping the walks at the j_{th} point.

$$\mathbf{F} = \mathbf{G}\mathbf{Y}$$

the probability of the i_{th} point which stops the random walks at the labeled data point whose label is j



In the special random walks, if an unlabeled sample **is similar to one of the labeled samples**, the walks starting from this unlabeled sample will **stop at the labeled sample with high probability.**

If this unlabeled **sample is an outlier or from a novel class**, this sample is not similar to any of the labeled samples, and the walks starting from this unlabeled sample will **stop at one of the unlabeled samples with high probability.**

Label Regression

$$\mathbf{y} = \mathbf{W}^T\mathbf{x} + \mathbf{b}$$

$$\boldsymbol{t}_i = [\underbrace{0, \ldots, 0}_{i-1}, 1, \underbrace{0, \ldots, 0}_{c-i}]^T$$

$$\mathbf{W} = \arg\min \mathcal{J}(\mathbf{W}, \mathbf{b})$$

$$\mathcal{J}(\mathbf{W}, \boldsymbol{b}) = \gamma\|\mathbf{W}\|^2 + \sum_{i=1}^{n}\sum_{j=1}^{c}\mathbf{F}_{ij}\|\mathbf{W}^T\boldsymbol{x}_i + \boldsymbol{b} - \boldsymbol{t}_j\|^2$$

Adaptive Semi-Supervised Learning

$$\left(\begin{array}{c} \min\limits_{W,b,Y} \left\|X_l^T W + \mathbf{1}_{nl} b^T - Y_l\right\|_F^2 + \sum\limits_{i=1}^{n}\sum\limits_{k=1}^{c} y_{ik}^r \left\|x_i^T W + b^T - t_k^T\right\|_F^2 \\ s.t.\ \ \forall i, y_{ik} \in [0,1], \sum\limits_{k=1}^{c} y_{ik} = 1 \end{array} \right) \Longrightarrow$$

1-KDD 2014

For clearly classified points, y_{i1}, y_{i2} would be one large and one small, For boundary points, however, y_{i1}, y_{i2} would be more likely equal

$$y_{i1} = 0.9, y_{i2} = 0.1, y_{i1}^r = 0.81, y_{i2}^r = 0.01 \qquad y_{i1} = 0.5, y_{i2} = 0.5, y_{i1}^r = 0.25, y_{i2}^r = 0.25$$

2-KDD 2014

Classification via Label-Adapted Kernels

Definition 1. **The label-adapted weight matrix is defined as**

$$W^{\bar{\chi}}(x_i, x_j) = m_1\left(\frac{\rho_1(x_i, x_j)^2}{\sigma_1}\right) m_2\left(\frac{\rho_2(\bar{\chi}(x_i), \bar{\chi}(x_j))^2}{\sigma_2}\right),$$

In particular, the minimization of the quadratic energy relies on the assumption that data points that belong to different classes will have a low similarity weight. This smoothness assumption, which is used in semi-supervised classification tasks, is often both local and global.

However, in many real data sets this might not be the situation, as h may not be smooth with respect to W

Transductive Classification (TC) problem: given labeled data $(x_1, y_1), \ldots, (x_l, y_l)$ and unlabeled data $x_{l+1}, \ldots, x_{l+u}, y_i \in \{-1, +1\}, 1 \le i \le l$.The goal is to predict the class labels of the given unlabeled data.

$$\min_{\mathbf{f} \in \mathbb{R}^n} \mathbf{f}^T \mathbf{L} \mathbf{f} + (\mathbf{f} - \mathbf{y})^T \mathbf{C} (\mathbf{f} - \mathbf{y}) \qquad \mathbf{f} = (f_1, \ldots, f_n)^T \in \mathbb{R}^n \qquad \mathbf{L} = \mathbf{D} - \mathbf{W}$$

$\mathrm{C} \in \mathbb{R}^{n \times n}$ is a diagonal matrix with the i_{th} diagonal element c_i set as: $c_i = \alpha_l > 0 \;\; 1 \le i \le l \quad c_i = \alpha_u \ge 0 \;\; l+1 \le i \le l+u$

$$w_{ij} = \exp\left(- \frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2} \right)$$

The first term of the objective function evaluates **how smooth the prediction f is with respect to the data manifold** which is represented by the regularization matrix **L**. It reflects the prior knowledge that a good prediction should **have low variance along the data manifold**.

The second term of the objective function uses a quadratic loss function to **measure the fitting error** of the prediction **f** . It constrains **f** to be close to ***y***.

GRF: f_i must be strictly equal to y_i for $1 \leq i \leq l$ and there is no constraint on the unlabeled data.

LapRLS: imposes a soft constraint on the labeled data but no constraint on the unlabeled data.

4-AAAI 2010

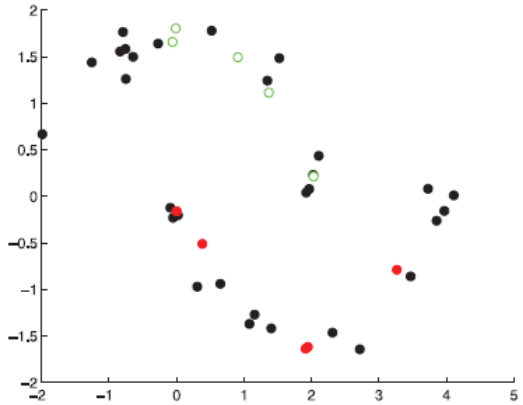
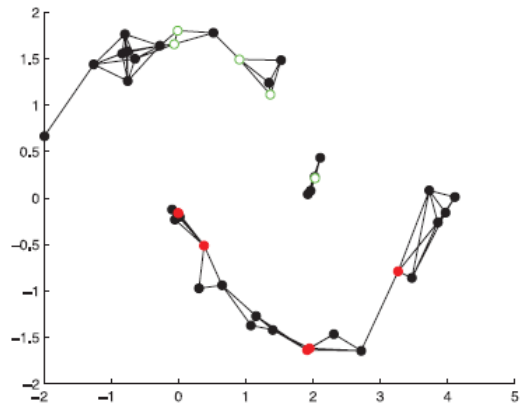
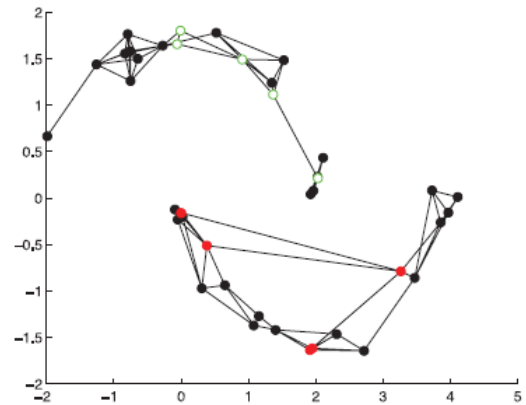
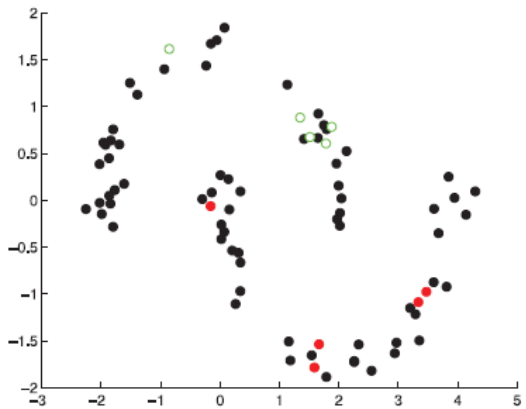
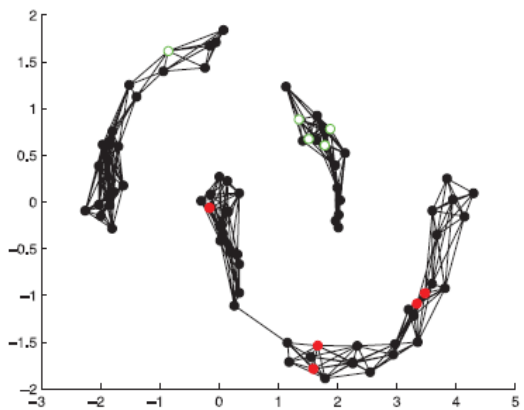
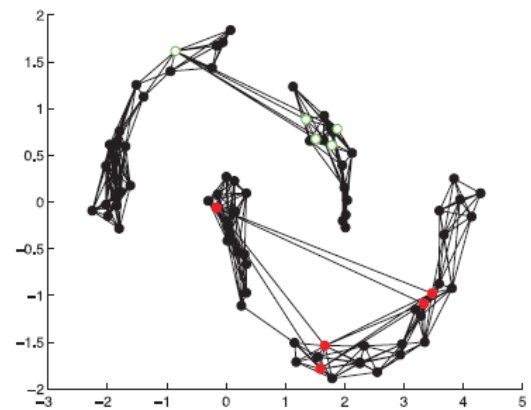
$$\min_{\mathbf{f}\in\mathbb{R}^n,\tilde{\mathbf{L}}\succeq 0} h(\mathbf{f},\tilde{\mathbf{L}}) = \mathbf{f}^T\tilde{\mathbf{L}}\mathbf{f} + (\mathbf{f}-\mathbf{y})^T\mathbf{C}(\mathbf{f}-\mathbf{y}) + \beta D(\tilde{\mathbf{L}},\mathbf{L})$$

D(A,B) is a measure of the dissimilarity between two matrices

$$\frac{\partial h}{\partial \tilde{\mathbf{L}}} = \mathbf{f}\mathbf{f}^T + \beta(\mathbf{L}^{-1} - \tilde{\mathbf{L}}^{-1}) = 0.$$

$$\frac{\partial h}{\partial \mathbf{f}_l} = 2(\tilde{\mathbf{L}}_{ll}\mathbf{f}_l + \tilde{\mathbf{L}}_{lu}\mathbf{f}_u + \alpha(\mathbf{f}_l - \mathbf{y}_l)) = 0$$

$$\frac{\partial h}{\partial \mathbf{f}_u} = 2(\tilde{\mathbf{L}}_{uu}\mathbf{f}_u + \tilde{\mathbf{L}}_{lu}^T\mathbf{f}_l) = 0.$$

JMLR 2006

LapRLS $$\min_{f\in\mathcal{H}_K}\frac{1}{l}\sum_{i=1}^{l}(y_i-f(x_i))^2+\gamma_A\|f\|_K^2+\frac{\gamma_I}{(u+l)^2}\mathbf{f}^T L\mathbf{f}$$

LapSVM $$\min_{f\in\mathcal{H}_K}\frac{1}{l}\sum_{i=1}^{l}(1-y_i f(x_i))_+ +\gamma_A\|f\|_K^2+\frac{\gamma_I}{(u+l)^2}\mathbf{f}^T L\mathbf{f}$$

6-IJCAI 2017

Linear Manifold Regularization with Adaptive Graph for Semi-supervised Dimensionality Reduction

$$\min_{W,b}\gamma_A\|W\|_F^2+\gamma_I Tr(W^T XLX^T W)+\frac{1}{l}\sum_{i=1}^{l}\|W^T x_i+b-y_i\|^2$$

Adaptive Neighbor Learning

$$\min_{s_i^T\mathbf{1}=1,0\le s_{ij}\le 1}\sum_{i,j=1}^{n}(\|x_i-x_j\|_2^2 s_{ij}+\gamma s_{ij}^2) \qquad Tr(W^T XLX^T W)=\frac{1}{2}\sum_{i,j=1}^{n}\|W^T x_i-W^T x_j\|_2^2 s_{ij}$$

$$\min_{W,b,S}\sum_{i,j=1}^{n}\|W^T x_i-W^T x_j\|_2^2 s_{ij}+\gamma\|S\|_F^2+\beta\|W\|_F^2+\alpha Tr(W^T X+b\mathbf{1}^T-Y)U(W^T X+b\mathbf{1}^T-Y)^T$$

$$s.t.\quad S\ge 0, S^T\mathbf{1}=\mathbf{1}$$

To **control the uniformity level of the manifold graph weights**, we use an entropy regularization term.

$$\min_{f,w_{ij}} \sum_{i,j=1}^{l+u} w_{ij}(f(x_i)-f(x_j))^2+\gamma_1 \sum_{i=1}^{l}(f(x_i)-y_i)^2+\gamma_2||f||_K^2+\eta\sum_{i,j=1}^{n} w_{ij}\ln w_{ij}$$

$$\text{s.t.}\sum_{j=1}^{u} w_{ij}=1, w_{ij}>0$$

In fact, each instance in the manifold graph is connected with only a few neighbor instances, thus only a few elements of each weight w_i should be non-zeros, and the rests should be zeros. That is, each vector weight vector w_i **should be sparse**.

$$\min_{f,w_{ij}} \sum_{i,j=1}^{l+u} w_{ij}(f(x_i)-f(x_j))^2+\gamma_1 \sum_{i=1}^{l}(f(x_i)-y_i)^2+\gamma_2||f||_K^2+\eta_1\sum_{i=1}^{u}(x_i-\sum_{i=1}^{u} w_{ij}x_j)^2+\eta_2\sum_{i=1}^{l+u}||w_{i\cdot}||_1$$

$$\text{s.t.}\sum_{j=1}^{u} w_{ij}=1, w_{ij}>0$$

Active Learning via Transductive Experimental Design 8-ICML 2006

Non-greedy Active Learning for Text Categorization using Convex Transductive Experimental Design 9-SIGIR 2008

Beyond the Point Cloud: from Transductive to Semi-supervised Learning 10-ICML 2005

Manifold Regularized Experimental Design for Active Learning 11-TIP 2017

Diversifying Convex Transductive Experimental Design for Active Learning 12-IJCAI 2016

2018.4.19

Experimental design

Classic experiment design considers learning a linear function $f(x) = w^T x \quad w \in \mathbb{R}^d$, from measurements $y_i = w^T x_i + \epsilon_i, i = 1, \dots, m, \epsilon_i \sim N(0, \sigma^2)$ is measurement noise. $x_1, .., x_m$ **are are experiments chosen from** $\boldsymbol{n}$ **candidates** $v_1, \dots, v_n \in \mathbb{R}^d, n > m$.

The goal of experimental design is to **find a set of experiments** x_i **that together are maximally informative**.

$$X: [x_1, \dots, x_m]^T \in \mathbb{R}^{m \times d}, set\{x_i\} \ |X| = m \quad \text{V}: [v_1, \dots, v_n]^T \in \mathbb{R}^{n \times d}, set\{v_i\} \ |V| = n$$

The maximum-likelihood estimate of w:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\{ J(\mathbf{w}) = \sum_{i=1}^{m} \left(\mathbf{w}^\top \mathbf{x}_i - y_i \right)^2 \right\}$$

Estimation error $e = w - \hat{w}$ mean: 0 covariance matrix: $\sigma^2 C_w$, C_w is the inverted Hessian of $J(w)$, σ is a constant

$$\mathbf{C_w} = \left(\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} = \left(\mathbf{X}^\top \mathbf{X} \right)^{-1}$$

The matrix C_w **characterizes the confidence of the estimation**, or **the informativeness of the selected data.**

A-optimal design

Let m_j denote the number of times for which v_j is chosen in **X,** $m_1 + \cdots m_n = m$

$$\min_{m_1,\ldots,m_n} \quad \mathrm{Tr}\left[\left(\sum_{j=1}^{n} m_j \mathbf{v}_j \mathbf{v}_j^\top \right)^{-1} \right]$$

$$\text{subject to} \quad m_j \geq 0, m_1 + \cdots + m_n = m, m_i \in \mathbb{Z}$$

$$\tau_j = m_j / m$$

$$\min_{\tau_1,\ldots,\tau_n} \quad \mathrm{Tr}\left[\left(\sum_{j=1}^{n} \tau_j \mathbf{v}_j \mathbf{v}_j^\top \right)^{-1} \right] \qquad \text{subject to} \quad \boldsymbol{\tau} \succeq 0, \mathbf{1}^\top \boldsymbol{\tau} = 1$$

Transductive Experimental Design

- The optimization criteria based on C_w **does not directly characterize the quality of predictions** on test data
- Standard experimental design **only considers linear functions** and is thus restrictive in applications.
- Very importantly, classic experimental design has to solve a SDP problem, which is often **very slow** when dealing with hundreds of data points

A general setting may consider a different set **T** of test data points besides candidates in **V**. Assume the two sets are the same.

$$\min_{\mathbf{w}} \left\{ J(\mathbf{w}) = \sum_{i=1}^{m} \left(\mathbf{w}^\top \mathbf{x}_i - y_i\right)^2 + \mu \|\mathbf{w}\|^2 \right\}$$

$$\mathbf{C}_\mathbf{w} = \left(\frac{\partial J(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^\top} \right)^{-1} = (\mathbf{X}^\top \mathbf{X} + \mu \mathbf{I})^{-1}$$

$\mathbf{f} = \left[f(v_1, \ldots, f(v_n))\right]$ be the function values on all the available data **V,**
the predictive error $\mathbf{f} - \hat{\mathbf{f}}$ has the covariance matrix $\sigma^2 C_{\mathrm{f}}$

$$\begin{aligned}\mathbf{C_f} &= \mathbf{V}\mathbf{C_w}\mathbf{V}^\top = \mathbf{V}(\mathbf{X}^\top\mathbf{X} + \mu\mathbf{I})^{-1}\mathbf{V}^\top \\ &= \frac{1}{\mu}\left[\mathbf{V}\mathbf{V}^\top - \mathbf{V}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mu\mathbf{I})^{-1}\mathbf{X}\mathbf{V}^\top\right]\end{aligned}$$

Woodbury matrix identity

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

The average predictive variance on **V** is given by $\frac{\sigma^2}{n}\mathrm{Tr}(C_{\mathrm{f}})$

Transductive experimental design

$$\begin{aligned}\max_{\mathbf{X}} \quad & Tr\left[\mathbf{V}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mu\mathbf{I})^{-1}\mathbf{X}\mathbf{V}^\top\right] \\ subject\ to \quad & \mathbf{X} \subset \mathbf{V}, |\mathbf{X}| = m\end{aligned}$$

$$\mathrm{Tr}(\mathbf{C_f}) = \mathrm{Tr}(\mathbf{C_w}\mathbf{V}^\top\mathbf{V})$$

Kernel Transductive Experimental Design

$$k(\mathbf{x}, \mathbf{v}) = \langle \phi(\mathbf{x}), \phi(\mathbf{v}) \rangle, \ \mathbf{x}, \mathbf{v} \in \mathbb{R}^d$$

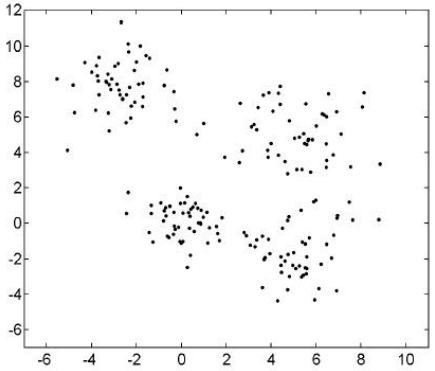
$$\begin{aligned} \max_{\mathbf{X}} \quad & \mathrm{Tr}\left[\mathbf{K_{vx}}(\mathbf{K_{xx}} + \mu\mathbf{I})^{-1}\mathbf{K_{xv}}\right] \\ \text{subject to} \quad & \mathbf{X} \subset \mathbf{V}, |\mathbf{X}| = m \end{aligned}$$

$$\begin{aligned} \max_{\mathbf{X}} \quad & Tr\left[\mathbf{V}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mu\mathbf{I})^{-1}\mathbf{X}\mathbf{V}^\top\right] \\ \textit{subject to} \quad & \mathbf{X} \subset \mathbf{V}, |\mathbf{X}| = m \end{aligned}$$

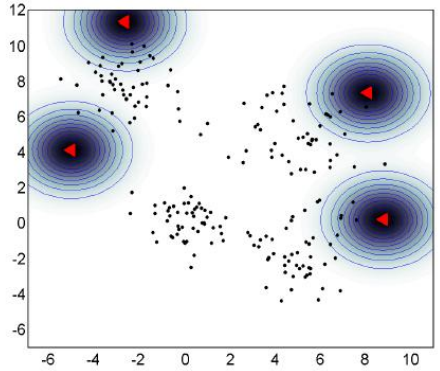
$$(\mathbf{K})_{ij} = k(\mathbf{v}_i, \mathbf{v}_j), \ (\mathbf{K_{vx}})_{ij} = k(\mathbf{v}_i, \mathbf{x}_j) \quad (\mathbf{K_{xx}})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$$

Sequential Design

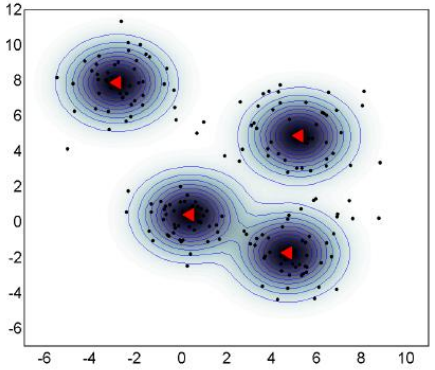
- Select $\mathbf{x} \in \mathbf{V}$ with the highest $\|\mathbf{K_x}\|^2/(k(\mathbf{x},\mathbf{x}) + \mu)$, and add $\mathbf{x}$ into $\mathbf{X}$, where $\mathbf{K_x}$ and $k(\mathbf{x},\mathbf{x})$ are $\mathbf{x}$'s corresponding column and diagonal entry in current $\mathbf{K}$;
- Update $\mathbf{K} \leftarrow \mathbf{K} - \frac{\mathbf{K_x}\mathbf{K_x}^\top}{(k(\mathbf{x},\mathbf{x})+\mu)}$



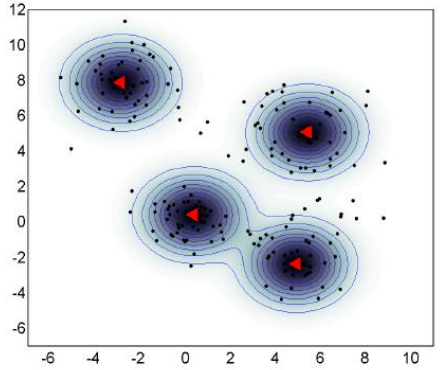
(a) Data set



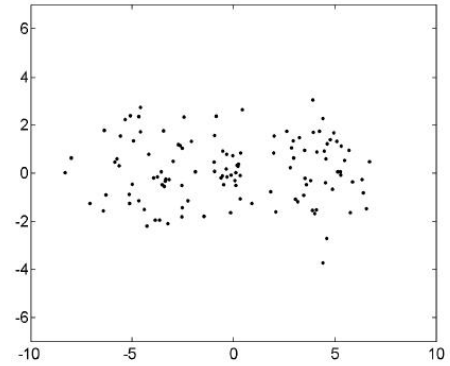
(b) A-optimal design



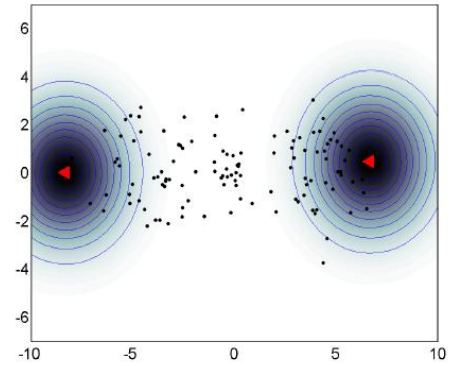
(c) Sequential design



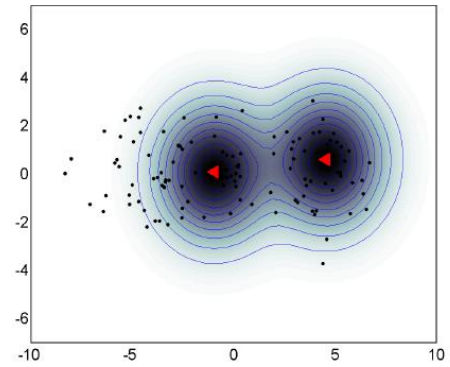
(d) Alternating design



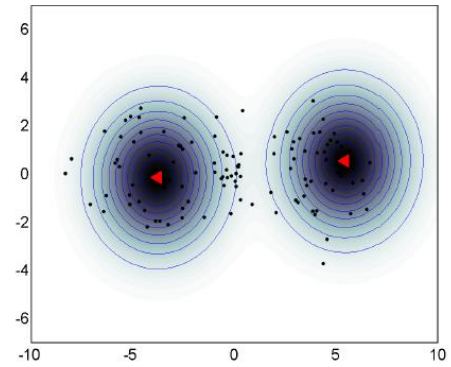
(a) Data set



(b) A-optimal design



(c) Sequential design



(d) Alternating design

$$\min_{\boldsymbol{\beta},\boldsymbol{\alpha}_i\in\mathbb{R}^N} \quad \sum_{i=1}^{M} \|\mathbf{x}_i - \mathbf{X}_{\mathcal{C}}^{\top}\mathbf{B}\boldsymbol{\alpha}_i\|^2 + \mu\|\mathbf{B}\boldsymbol{\alpha}_i\|^2 + \gamma\|\boldsymbol{\beta}\|_1$$
$$\text{subject to} \quad \mathbf{x}_i \in \mathbf{X}_{\mathcal{P}}, \quad \mathbf{B} = \text{diag}(\boldsymbol{\beta}), \quad \mathbf{B} \succeq 0.$$

9-SIGIR 2008

$$\min_{\boldsymbol{\beta},\boldsymbol{\alpha}_i\in\mathbb{R}^N} \quad \sum_{i=1}^{M} \|\mathbf{x}_i - \mathbf{X}_{\mathcal{C}}^{\top}\boldsymbol{\alpha}_i\|^2 + \sum_{j=1}^{N} \frac{\alpha_{i,j}^2}{\beta_j} + \gamma\|\boldsymbol{\beta}\|_1$$
$$\text{subject to} \quad \mathbf{x}_i \in \mathbf{X}_{\mathcal{P}}, \quad \beta_j \geq 0, \quad j = 1, \cdots, N,$$

$$\boldsymbol{\beta} = [\beta_1, \ldots, \beta_N]$$
$$\boldsymbol{\alpha}_i = [\alpha_{i,1}, \ldots, \alpha_{i,N}]^{\top}$$

$$\frac{\sum_{i=1}^{M} \alpha_{i,j}^2}{|\beta_j|} + \gamma|\beta_j| \geq 2\sqrt{\gamma \sum_{i=1}^{M} \alpha_{i,j}^2},$$

$$\beta_j^2 = \frac{1}{\gamma}\sum_{i=1}^{M} \alpha_{i,j}^2$$

$$\min_{\boldsymbol{\alpha}_i\in\mathbb{R}^N} \quad \sum_{i=1}^{M} \|\mathbf{x}_i - \mathbf{X}_{\mathcal{C}}^{\top}\boldsymbol{\alpha}_i\|^2 + 2\sum_{j=1}^{N} \sqrt{\gamma \sum_{i=1}^{M} \alpha_{i,j}^2}$$

11-TIP 2017

Find the informative samples by minimizing the expected prediction variance on the test data.

$$\tilde{k}(x,z) = k(x,z) - \mathbf{k}_x^t (I + MK)^{-1} M \mathbf{k}_z$$

10-ICML 2005

$$\min_{\alpha_i \in R^l} \sum_{i=1}^{l} \left\| \tilde{\phi}(x_i) - \tilde{\phi}(Z)\alpha_i \right\|^2 + \gamma_1 \|\alpha_i\|^2$$

$$\tilde{\phi}(Z) = [\tilde{\phi}(z_1), \dots, \tilde{\phi}(z_l)]$$

$$\min_{\alpha_i, \beta \in R^l} \sum_{i=1}^{l} \left(\left\| \tilde{\phi}(x_i) - \tilde{\phi}(X)\alpha_i \right\|^2 + \sum_{j=1}^{n} \frac{\alpha_{i,j}^2}{\beta_j} \right) + \lambda \|\beta\|_1$$

$$s.t. \beta_j \geq 0, \qquad j = 1, \dots, n$$

$$\alpha_i = (\alpha_{i,1}, \dots, \alpha_{i,n})^T$$

分割线

$$\min_{\mathcal{A}, \boldsymbol{\alpha}_i \in \mathbb{R}^K} \sum_{i=1}^{M} \|\mathbf{x}_i - \mathbf{X}_{\mathcal{A}}^{\top} \boldsymbol{\alpha}_i\|^2 + \mu \|\boldsymbol{\alpha}_i\|^2$$

$$\text{subject to} \quad |\mathcal{A}| = K, \quad \mathcal{A} \subset \mathcal{C}, \quad \mathbf{x}_i \in \mathbf{X}_{\mathcal{P}}$$

$$\min_{\boldsymbol{\beta}, \boldsymbol{\alpha}_i \in \mathbb{R}^N} \sum_{i=1}^{M} \|\mathbf{x}_i - \mathbf{X}_{\mathcal{C}}^{\top} \boldsymbol{\alpha}_i\|^2 + \sum_{j=1}^{N} \frac{\alpha_{i,j}^2}{\beta_j} + \gamma \|\boldsymbol{\beta}\|_1$$

$$\text{subject to} \quad \mathbf{x}_i \in \mathbf{X}_{\mathcal{P}}, \quad \beta_j \geq 0, \quad j = 1, \cdots, N,$$

$$\boldsymbol{\beta} = [\beta_1, \dots, \beta_N]$$

$$\boldsymbol{\alpha}_i = [\alpha_{i,1}, \dots, \alpha_{i,N}]^{\top}$$

9-SIGIR 2008

11-TIP 2017

Algorithm 2 Convex TED

Require: candidates $\mathbf{X}_{\mathcal{C}}$, unlabeled data $\mathbf{X}_{\mathcal{P}}$, $\gamma > 0$;

1: initialize $(\alpha_{i,j})$;
2: **repeat**
3: $\beta_j \leftarrow \sqrt{\frac{1}{\gamma}\sum_{i=1}^{M}\alpha_{i,j}^2}$ for $j = 1, \ldots, N$;
4: $\boldsymbol{\alpha}_i \leftarrow (\mathrm{diag}(\boldsymbol{\beta})^{-1} + \mathbf{X}_{\mathcal{C}}\mathbf{X}_{\mathcal{C}}^{\top})^{-1}\mathbf{X}_{\mathcal{C}}\mathbf{x}_i$, for $i = 1, \ldots, M$;
5: **until** converge;
6: $\mathbf{X}_{\mathcal{A}} \leftarrow \{\mathbf{x}_j | \mathbf{x}_j \in \mathbf{X}_{\mathcal{C}}, \beta_j \neq 0\}$;
7: **return** $\mathbf{X}_{\mathcal{A}}$

Algorithm 1 MRED for Active Learning

Input: The n unlabeled data samples X, the number of the selected most information data samples l, the number of the nearest neighbor data samples k

Step 1: Construct a nearest neighbor Laplacian graph with the weight matrix W as calculated in Eq. (15) on the unlabeled samples X and calculate

Step 2: Construct the kernel Gram matrix K with an selected input kernel type and let $M = L$.

Step 3: Construct the data-dependent deformed kernel Gram matrix $\tilde{K}$ according to Eq. (14).

Step 4: Let u_i be the ith column vector of K and initialize $\alpha_{i,j} = 1$.

Step 4.1: Repeat

Step 4.2: Compute β_j according to Eq. (29), i.e., $\beta_j = \sqrt{\sum_{i=1}^{n}\alpha_{i,j}^2/\lambda}$.

Step 4.3: Compute α_{i} according to Eq. (27), i.e., $\alpha_{\mathrm{i}} = (D_{\beta}^{-1} + \tilde{K})^{-1}\tilde{K}_i$.

Step 4.4: Until Convergence

Step 5: Rank the samples in X by following $\beta_j (j = 1, \ldots, n)$ in a descending order and then return the top l samples as the selected most informative ones Z.

Output The l selected most informative samples can be labeled as the training samples.

Manifold Regularization

Adaptive Graph

Transductive Experimental Design

Active Learning

$$\min_{w,L}\left\{J(w)=\sum_{i=1}^{m}(w^T x_i-y_i)^2+\mu||w||^2+\gamma \mathrm{f}^T L\mathrm{f}\right\}$$

$$\min_{w,L}\ \{J(w)=(X^T\mathrm{w}-\mathrm{y})^2+\mu||w||^2+\gamma \mathrm{w}^T XLX^T\mathrm{w}\}$$

$$\min_{\mathbf{w}}\left\{J(\mathbf{w})=\sum_{i=1}^{m}\left(\mathbf{w}^\top\mathbf{x}_i-y_i\right)^2+\mu\|\mathbf{w}\|^2\right\}$$

$$\mathbf{C}_{\mathbf{w}}=\left(\frac{\partial J(\mathbf{w})}{\partial\mathbf{w}\partial\mathbf{w}^\top}\right)^{-1}=(\mathbf{X}^\top\mathbf{X}+\mu\mathbf{I})^{-1}$$

$$\mathbf{C}_{\mathbf{f}}=\mathbf{V}\mathbf{C}_{\mathbf{w}}\mathbf{V}^\top$$

Diversifying Convex Transductive Experimental Design for Active Learning

12-IJCAI 2016

$$\min_{\mathbf{A},\mathbf{b}} \quad ||\mathbf{X}-\mathbf{X}\mathbf{A}||_F^2+\sum_{i=1}^{n}\frac{\sum_{j=1}^{n}a_{ij}^2}{b_i}+\gamma||\mathbf{b}||_1+\alpha\mathbf{b}^T\mathbf{S}\mathbf{b}$$
$$s.t. \quad b_i\geq 0, i=1,2,...,n$$

9-SIGIR 2008

$$\min_{\boldsymbol{\beta},\boldsymbol{\alpha}_i\in\mathbb{R}^N} \quad \sum_{i=1}^{M}\|\mathbf{x}_i-\mathbf{X}_{\mathcal{C}}^{\top}\boldsymbol{\alpha}_i\|^2+\sum_{j=1}^{N}\frac{\alpha_{i,j}^2}{\beta_j}+\gamma\|\boldsymbol{\beta}\|_1$$
$$\text{subject to} \quad \mathbf{x}_i\in\mathbf{X}_{\mathcal{P}}, \quad \beta_j\geq 0, \quad j=1,\cdots,N,$$

This constraint guarantees that highly similar samples would not have higher scores in sample selection at the same time.